

NELL2RDF: Reading the Web, Tracking the Provenance, and Publishing It as Linked Data

José M. Giménez-García¹, Maísa Duarte¹, Antoine Zimmermann²
Christophe Gravier¹, Estevam R. Hruschka Jr.^{3,4}, and Pierre Maret¹

¹ UJM-Saint-Étienne, Laboratoire Hubert Curien, Saint Étienne, France
`{jose.gimenez.garcia,maisa.duarte,christophe.gravier,
pierre.maret}@univ-st-etienne.fr`

² MINES Saint-Étienne, Laboratoire Hubert Curien, Saint-Étienne, France
`antoine.zimmermann@emse.fr`

³ Federal University of Sao Carlos - UFSCar, São Carlos, Brazil

⁴ Carnegie Mellon University - CMU, Pittsburgh, United States
`estevam@cs.cmu.edu`

Abstract. NELL is a system that continuously reads the Web to extract knowledge in the form of entities and relations between them. It has been running since January 2010 and extracted over 450 million candidate statements, 28 million of which remain in iteration 1100. NELL’s generated data comprises all the candidate statements, together with detailed metadata information about how it was generated. This information includes how each component of the system contributed to the extraction of the statement, as well as when that happened and how confident the system is in the veracity of the statement. However, the data is only available in an ad hoc CSV format that makes it difficult to exploit out of the context of NELL. In order to make it more usable for other communities, we adopt Linked Data principles to publish a more standardized, self-describing dataset with rich provenance metadata.

Keywords: NELL·RDF·Metadata·Reification·Provenance

1 Introduction

Never-Ending Language Learning (NELL) [2] is an autonomous computational system with the aim of learning continually and incrementally. It generates a knowledge base where beliefs are learned from the Web using an ontology previously created to guide the learning. One of the most significant resource contributions of NELL is the metadata attached to each one of the millions of beliefs. This consists of provenance data about how categories, relations and concepts are extracted, and the confidence about the process itself. It evolves in every iteration, and is used by NELL to continuously retrain NELL’s learning components, in order to improve its understanding about what it reads from the Web. NELL runs sequential iterations. In each of them, new candidate beliefs can be created, older candidates can be promoted to a status of higher credibility, promoted beliefs can be demoted again, or beliefs can be discarded altogether. NELL has been running for 8 years in Carnegie Mellon University. In

total, over the years, NELL has collected over 450 million candidate beliefs and promoted 10 million of them. In iteration 1100, 28 million candidate beliefs and 2.6 million promoted beliefs remain. Zimmermann et al. [26] made a modest attempt to convert NELL’s beliefs and ontology into RDF and OWL, but their work completely disregarded candidate beliefs and metadata, including less than 0.04% of the data of the current work. Thus, we redesign the dataset to include all the provenance metadata for all beliefs—candidate or promoted—and fully automatize the processing of new iterations so that we can guarantee its sustainability. Furthermore, metadata about beliefs require a reification model for which there are several representations. We publish variants of the dataset according to well-known reification models in the community so it can be more easily applied in a wider set of scenarios.

This work fills several important gaps: (1) It can be used as a general knowledge base with millions of statements annotated with varying degrees of confidence. (2) As a valuable resource for research in managing and exploiting meta-knowledge. (3) It exposes the Never-Ending Learning community in general, and NELL in particular, to the potential of Linked data, allowing to connect their research results using Linked data principles. (4) It helps understanding NELL’s metadata by structuring and self-documenting the output of its components.

The rest of the paper is organized as follows: Sec. 2 presents related work; Section 3 describes the transformation of NELL’s data and metadata to RDF and how it is published; finally, Section 4 provides final remarks and future work.

2 Related Work

This section describes existing works in the Semantic Web community that relates to NELL, as well as research in representing metadata about statements using Semantic Web technologies.

2.1 NELL and the Semantic Web

A first experiment in translating NELL’s data to RDF was made in 2013 [26]. Only the promoted beliefs were considered, and no metadata about the provenance of belief was generated. This resulted in an RDF dataset with 5.8 million triples (less than 0.04% of our current data) providing information about 1.5 million entities. Among the 2.5 million distinct object values, 99% were literals associating labels to entities, and most of the remaining triples were assigning `rdf:type` to the entities. Less than 1% of the triples were relations between instances in this data set. In spite of these strong limitations, the NELL2RDF dataset was exploited by a few research works that analyze and enhance data quality [24, 9]. Moreover, NELL has proven useful for some tasks in Semantic Web research, such as improving precision of relation extraction [21], type prediction [18, 19], or alignment with DBpedia [6, 25], but these works only used a very small portion of NELL’s data. Other papers, while citing NELL as a prominent example of open knowledge graph, do not make any use of its data. As noted by Gerber et al. [11], NELL’s data cannot be directly integrated in the Web of Data. This research would benefit from having a formal representation in linked data of NELL.

2.2 Statements about Statements in the Semantic Web

The RDF data model only allows to represent binary (or dyadic) information. That is, a single relation between two entities. However, it is sometimes necessary to express additional information about the statements themselves. For that reason, a number of approaches have sprung in the recent years: RDF reification [1, Sec. 5.3] represents the statement using a resource, and then creates triples to indicate the subject, predicate and object of the statement; N-Ary relations [23] create a new resource that identifies the relation and connects subject and object using different design patterns; named graphs [3] add a fourth element to each triple, that can be used to identify a triple or set of triples later on; the Singleton Property [22] creates a unique property for each triple, related to the original property; and NdFluents [13] creates a unique version of the subject and the object (in the case it is not a literal) of the triple, and attaches them to the original resources and the context of the statement. Wikidata makes use of N-Ary relations [7], while Nano-publications use named graphs [20].

Some works [14, 15, 10] compare a number of reification approaches, although the size of the datasets they use is relatively modest (an old Wikidata set, with around 81 million triples, and a small subset of DBpedia to which some revision history data is attached, with 1 billion triples approximately). These experiments yield non-conclusive results about which representation is optimal. Hence, in order to make NELL2RDF more easily applicable in a wider set of scenarios, we provide datasets in all different approaches.

3 NELL2RDF

NELL’s beliefs are published in tab-separated format, where each line contains a number of fields to express the belief and the associated metadata, such as iteration of promotion, confidence score, or the activity of the components that inferred the belief. Each line is converted into a triple representing the belief, plus additional triples containing the types and all the associated labels for subject and object, as well as a preferred label using the `skos:prefLabel` property. Then, each belief is reified into a resource, to which the provenance metadata is attached. We provide for each iteration five different datasets with different reification approaches, namely RDF reification [1, Sec. 5.3], N-Ary relations [23], named graphs [3], singleton Properties [22], and NdFluents [13], as well as the dataset without annotations.

The ontology can be seen in Figure 1. We make use of the PROV-O ontology [17] to describe the provenance. Each `Belief` can be related with one or more `ComponentExecution` that, in turn, are performed by a `Component`. If the belief is a `PromotedBelief`, it has attached its `iterationOfPromotion` and `probabilityOfBelief`. The `ComponentIteration` is related to information about the process: the `iteration`, `probabilityOfBelief`, `Token`, `source` and `atTime` (the date and time it was processed). The `Token` expresses the concepts that the `Component` is relating together. Those concepts can be a pair of entities for a `RelationToken`, and an entity and a class for a `GeneralizationToken` (note that `LatLong` component has a different token `GeoToken`, further described later). Finally, each component has a `source` string describing their process for the belief. This string is then further analyzed and

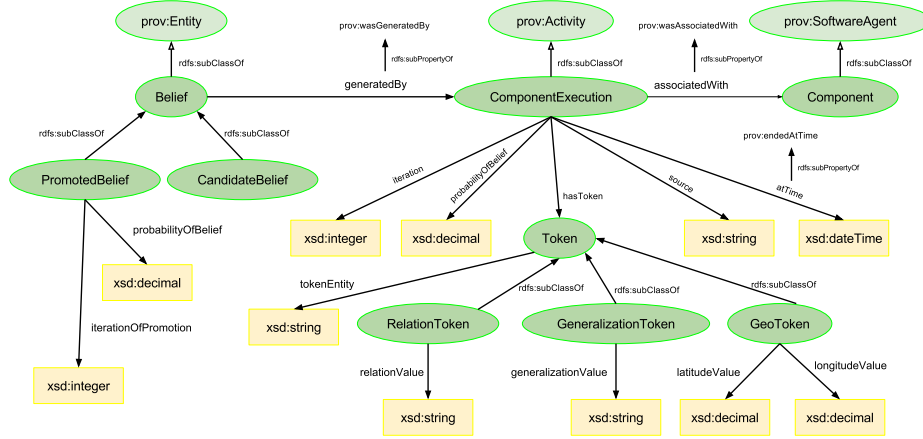


Fig. 1. NELL2RDF metadata ontology

translated into a different set of IRIs for each type of components. We describe in the web page⁵ the classes and properties related to each component of the system.

The current version of NELL2RDF includes promoted and candidate beliefs of iterations 1075, 1090, and 1100, as well as the provenance for all beliefs, for a total of more than 14.5 billion triples. It also contains the ontologies for the beliefs and the provenance metadata. Metadata about the dataset is modeled using VoID and DCAT vocabularies. The results indicate that, when the amount of metadata per statement is significant, the size both in bytes and in number of triples is similar for any reification approach. The model can affect, however, the efficiency of compressed serializations or indexes: The size of the singleton property dataset is 30% than the rest in HDT format, due to having a big number of different properties.

In addition, NELL2RDF entities are linked to DBPedia (296255 in iteration 1100), generated using the beliefs about the Wikipedia pages of NELL entities. While only a first step to interlink NELL2RDF to the linked data cloud, it increases its usability and shows the potential for further research in this aspect.

NELL2RDF is available at the canonical URL <http://w3id.org/nellrdf>, where we provide the datasets in gzipped N-Triples and HDT [8] format, as well as the SPARQL endpoints for each model and for the dataset without metadata. Due to the sheer size of the datasets, the HDT generation was performed using HDT-MR [12]. The datasets and all related information are published under the *Creative Commons CC0 1.0 Universal* license⁶.

4 Discussion and Future Work

In this work we present NELL2RDF and make it available to the research community as a reference dataset of general knowledge, containing statement-level provenance

⁵ <http://w3id.org/nellrdf>

⁶ <https://creativecommons.org/publicdomain/zero/1.0/legalcode>

metadata and confidence scores given by NELL’s inner works. We hope to bridge the gap between the NELL community and the Semantic Web community, drawing attention from the former to Linked Data standards and practices, and to the latter to the never-ending learning paradigms. We believe that NELL2RDF presents a lot of potential use cases for research: Alignment with other knowledge bases could help to add information or resource lexicalizations, or improve the accuracy of their statements. While we already made a step in that direction, but we think there is a lot of potential research to be done. NELL2RDF contains a big proportion of metadata statements, encoded using five different reification approaches; this makes of it an ideal testbed to compare how different metadata representations behave (in a similar fashion as Hernández et al. [14, 15] and Frey et al. [10]). Existing research shows that there is interest in NELL for tasks like relation extraction [21], type prediction [18, 19], or quality analysis [24, 9], but its usage has not taken off. NELL2RDF will enable further research in this direction. It can also be exploited as an additional resource for comparison against new research in relation extraction, or tasks such as entity disambiguation or query answering. In addition, NELL is starting to be explored in languages different than English, such as Portuguese [16, 4] and French [5]. Our intention is to convert those datasets to RDF as they become available to the public. This will also allow to explore mappings between languages in NELL. Finally, we are processing NELL’s historical data and adding older iterations, and we plan to merge all data in a unique contextualized knowledge graph that uses iteration and provenance as two different contexts. This will allow to query and explore how data has evolved over time and what new information was the cause of changes.

Acknowledgements: This work is supported by H2020 Marie Skłodowska-Curie ITN No 642795. We would like to thank Bryan Kisiel from NELL’s CMU and Thomas Gautrais from Laboratoire Hubert Curien, Saint Etienne, for their technical support.

References

- [1] Brickley, D., Guha, R.: RDF Schema 1.1. W3C Recommendation (2014).
- [2] Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E.R., Mitchell, T.M.: Toward an Architecture for Never-Ending Language Learning. AAAI (2010).
- [3] Carroll, J.J., Bizer, C., Hayes, P.J., Stickler, P.: Named graphs. J. Web Sem. (2005).
- [4] Duarte, M.C., Hruschka, E.R.: How to Read The Web In Portuguese Using the Never-Ending Language Learner’s Principles. ISDA (2014).
- [5] Duarte, M.C., Maret, P.: Vers une instance française de NELL : chaîne TLN multilingue et modélisation d’ontologie. EGC (2017).
- [6] Dutta, A., Meilicke, C., Ponzetto, S.P.: A Probabilistic Approach for Integrating Heterogeneous Knowledge Sources. ESWC (2014).
- [7] Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandečić, D.: Introducing Wikidata to the Linked Data Web. ISWC (2014).
- [8] Fernández, J.D., Martínez-Prieto, M.A., Gutiérrez, C., Polleres, A., Arias, M.: Binary RDF representation for publication and exchange (HDT). JWS (2013).
- [9] Fleischhacker, D., Paulheim, H., Bryl, V., Völker, J., Bizer, C.: Detecting Errors in Numerical Linked Data Using Cross-Checked Outlier Detection. ISWC (2014).

- [10] Frey, J., Müller, K., Hellmann, S., Rahm, E., Vidal, M.-E.: Evaluation of Metadata Representations in RDF stores. *Sem. Web. J.* (2017).
- [11] Gerber, D., Ngomo, A.-C.N.: Extracting Multilingual Natural-Language Patterns for RDF Predicates. *EKAW* (2012).
- [12] Giménez-García, J.M., Fernández, J.D., Martínez-Prieto, M.A.: HDT-MR: A Scalable Solution for RDF Compression with HDT and MapReduce. *ESWC* (2015).
- [13] Giménez-García, J.M., Zimmermann, A., Maret, P.: NdFluents: An Ontology for Annotated Statements with Inference Preservation. *ESWC* (2017).
- [14] Hernández, D., Hogan, A., Krötzsch, M.: Reifying RDF: What Works Well With Wikidata? *SSWS* (2015).
- [15] Hernández, D., Hogan, A., Riveros, C., Rojas, C., Zerega, E.: Querying Wikidata: Comparing SPARQL, Relational and Graph Databases. *ISWC* (2016).
- [16] Hruschka, E.R., Duarte, M.C., Nicoletti, M.C.: Coupling as Strategy for Reducing Concept-Drift in Never-Ending Learning Environments. *Fund. Inform.* (2013).
- [17] Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: PROV-O: The PROV Ontology. *W3C Recommendation* (2013).
- [18] Melo, A., Paulheim, H., Völker, J.: Type Prediction in RDF Knowledge Bases Using Hierarchical Multilabel Classification. *WIMS* (2016).
- [19] Melo, A., Völker, J., Paulheim, H.: Type Prediction in Noisy RDF Knowledge Bases Using Hierarchical Multilabel Classification with Graph and Latent Features. *Int. J. Artif. Intell. Tools* (2017).
- [20] Mons, B., Velterop, J.: Nano-Publication in the e-science era. *SWASD* (2009).
- [21] Moro, A., Li, H., Krause, S., Xu, F., Navigli, R., Uszkoreit, H.: Semantic Rule Filtering for Web-Scale Relation Extraction. *ISWC* (2013).
- [22] Nguyen, V., Bodenreider, O., Sheth, A.: Don't like RDF Reification?: Making Statements about Statements Using Singleton Property. *WWW* (2014).
- [23] Noy, N., Rector, A., Hayes, P., Welty, C.: Defining N-Ary Relations on the Semantic Web. *W3C Working Group* (2006).
- [24] Paulheim, H., Bizer, C.: Improving the Quality of Linked Data Using Statistical Distributions. *Int. J. Semantic Web Inf. Syst.* (2014).
- [25] Wijaya, D.T., Mitchell, T.M.: Mapping Verbs in Different Languages to Knowledge Base Relations using Web Text as Interlingua. *NAACL-HLT* (2016).
- [26] Zimmermann, A., Gravier, C., Subercaze, J., Cruzille, Q.: Nell2RDF: Read the Web, and Turn it into RDF. *KNOW@LOD* (2013).